

# Analisis Segmentasi Pelanggan Menggunakan Kombinasi RFM Model dan Teknik *Clustering*

Beta Estri Adiana<sup>1</sup>, Indah Soesanti<sup>2</sup>, Adhistya Erna Permanasari<sup>3</sup>  
 Jurusan Teknik Elektro dan Teknologi Informasi, Universitas Gadjah Mada  
 Jl. Grafika No.2, Yogyakarta  
[beta.estri.a@mail.ugm.ac.id](mailto:beta.estri.a@mail.ugm.ac.id)  
 Jurusan Teknik Elektro dan Teknologi Informasi, Universitas Gadjah Mada  
 Jl. Grafika No.2, Yogyakarta  
[indsanti@gmail.com](mailto:indsanti@gmail.com)  
 Jurusan Teknik Elektro dan Teknologi Informasi, Universitas Gadjah Mada  
 Jl. Grafika No.2, Yogyakarta  
[adhistya@ugm.ac.id](mailto:adhistya@ugm.ac.id)

**Abstract** – Intense competition in the business field motivates a small and medium enterprises (SMEs) to manage customer services to the maximal. Improve of customer royalty by grouping customers into some of groups and determining appropriate and effective marketing strategies for each group. Customer segmentation can be performed by data mining approach with clustering method. The main purpose of this paper is customer segmentation and measure their loyalty to a SME's product. Using CRISP-DM method which consist of six phases, namely business understanding, data understanding, data preparation, modeling, evaluation and deployment. The K-Means algorithm is used for cluster formation and RapidMiner as a tool used to evaluate the result of clusters. Cluster formation is based on RFM (recency, frequency, monetary) analysis. Davies Bouldin Index (DBI) is used to find the optimal number of clusters (k). The customers are divided into 3 clusters, total of customer in first cluster is 30 customers who entered in typical customer category, the second cluster there are 8 customers who entered in superstar customer and 89 customers in third cluster is dormant cluster category.

**Keywords** – customer segmentation, data mining, clustering, CRISP-DM, RFM model, K-Means algorithm.

**Intisari** – Persaingan yang ketat di bidang bisnis memotivasi sebuah usaha kecil dan menengah (UKM) untuk mengelola pelayanan terhadap konsumen tetap (pelanggan) secara maksimal. Meningkatkan kesetiaan pelanggan dengan mengelompokkan pelanggan menjadi beberapa kelompok dan menentukan strategi pemasaran yang tepat dan efektif untuk setiap kelompok. Segmentasi pelanggan dapat dilakukan melalui pendekatan *data mining* dengan metode *clustering*. Tujuan utamanya adalah segmentasi pelanggan dan mengukur kesetiaan mereka terhadap sebuah produk UKM. Dengan menggunakan metode CRISP-DM yang terdiri dari enam fase, yakni pemahaman bisnis (*business understanding*), pemahaman data (*data understanding*), persiapan data (*data preparation*), pemodelan (*modeling*), evaluasi (*evaluation*), dan penerapan (*deployment*). Algoritma K-means digunakan untuk pembentukan klaster dan RapidMiner sebagai *tool* yang digunakan untuk mengevaluasi klaster-klaster yang terbentuk. Pembentukan klaster didasarkan pada analisa RFM ( *Recency*,

*Frequency*, dan *Monetary*). *Davies Bouldin Index* (DBI) digunakan untuk menemukan jumlah *cluster* (k) yang optimal. Hasilnya kelompok pelanggan yang terbentuk ada tiga kelompok dengan kelompok pertama berjumlah 30 pelanggan masuk dalam kategori *typical customer*, kelompok kedua terdapat 8 pelanggan yang masuk dalam kategori *superstar customer* dan kelompok ketiga berjumlah 89 pelanggan pada kategori *dormant customer*.

**Kata Kunci** – segmentasi pelanggan, *data mining*, *clustering*, CRISP-DM, model RFM, algoritma K-Means.

## I. PENDAHULUAN

Sejalan dengan peradaban manusia mempengaruhi perkembangan teknologi informasi (TI). Perkembangan TI mempengaruhi beberapa sektor, seperti di bidang bisnis, bidang kesehatan, bidang pendidikan, dan juga pemerintahan. Dalam persaingan bisnis, perusahaan dituntut untuk dapat memanfaatkan kemampuan yang ada semaksimal mungkin agar mampu bersaing dengan perusahaan lain.

Perubahan kebutuhan pelanggan mendorong terjadinya perubahan dalam bidang pemasaran. Bagian pemasaran perusahaan memiliki peran penting dalam menghadapi persaingan yang semakin ketat dengan perusahaan lain. Perusahaan yang berorientasi pasar atau dengan kata lain menjual produk ke pelanggan, umumnya akan menghadapi masalah di bidang pemasaran. Diperlukan survei pasar untuk mendapatkan informasi tentang permintaan dan kebutuhan spesifik pelanggan [1]. Perusahaan harus mempertimbangkan karakteristik setiap pelanggan. Informasi itu digunakan untuk mengembangkan produk yang mengandung kombinasi dan atribut optimal yang diinginkan oleh pelanggan.

Saat ini, perusahaan modern tidak hanya fokus pada strategi yang mengutamakan produk (*product/service oriented*), mereka juga melakukan strategi yang

mengutamakan pelanggan (*customer oriented*). Hal ini penting dilakukan untuk perkembangan perusahaan dengan menciptakan hubungan baik antara perusahaan dan pelanggan [2][3]. Situasi ekonomi masyarakat yang semakin berubah, dengan persaingan industri pangan yang semakin ketat, ditandai dengan meningkatnya jumlah produk serupa yang ditawarkan kepada pelanggan. Teknologi informasi menjadi sangat berpengaruh bagi kemajuan sebuah industri. Teknologi informasi membantu memberikan layanan atau produk yang lebih baik kepada pelanggan [4].

UD Gemilang Kencana merupakan salah satu usaha kecil dan menengah (UKM) di Kabupaten Wonosobo yang bergerak di bidang industri makanan pengolahan buah Carica menjadi manisan. Industri ini menyadari akan pentingnya hubungan antara pelanggan yang loyal dengan keberhasilan bisnis. UD Gemilang Kencana setiap harinya harus memenuhi kebutuhan konsumen dan dituntut untuk dapat mengambil keputusan yang tepat dalam menentukan strategi penjualan. Konsumen yang datang dari berbagai kalangan dengan selera yang berbeda, menuntut industri ini harus mampu menarik minat beli konsumen. Dengan jumlah konsumen yang tergolong banyak dan terus meningkat mengakibatkan UD Gemilang Kencana kesulitan mengidentifikasi minat beli konsumen dan selera konsumen terhadap produknya. Penggunaan metode manual atau tradisional untuk melakukan identifikasi konsumen dari data membutuhkan kemampuan manusia untuk menganalisis dan menginterpretasikan data. Dengan volume data yang berkembang, baik dari jumlah *record* dan jumlah *field*. Hal ini dapat mengakibatkan UD Gemilang Kencana kehilangan konsumen potensial dan merugikan perusahaan.

Segmentasi pelanggan diperlukan untuk mengelompokkan pelanggan yang memiliki kesamaan karakteristik. Hal ini diperlukan untuk mengetahui perilaku (*behavior*) konsumen sehingga akan membantu dalam penerapan strategi pemasaran (*marketing*) yang tepat untuk meningkatkan pendapatan perusahaan. Penggunaan teknik *data mining* merupakan salah satu solusi untuk persoalan segmentasi konsumen. *Data mining* diharapkan dapat membantu proses pengambilan keputusan yang tepat, memungkinkan perusahaan atau industri untuk mengelola data yang tersimpan dalam suatu basis data, *data warehouse*, atau tempat penyimpanan lainnya menjadi sebuah informasi dan pengetahuan (*knowledge*) yang baru. *Data mining* dapat mengekstrak pengetahuan berharga dari sejumlah data yang diatur oleh pemahaman manusia [5] sehingga pengembangan aplikasi *data mining* penting untuk pengembangan bisnis.

*Clustering* merupakan teknik *data mining* yang membagi data dalam suatu himpunan ke dalam beberapa kelompok yang mana kesamaan data dalam suatu kelompok lebih besar dibandingkan kesamaan data tersebut dengan

data dalam kelompok lain [6]. Semakin kecil jumlah *cluster* yang digunakan, maka akan mempermudah pemahaman terhadap struktur data di dalamnya, tetapi pola informasi penting yang ada akan terabaikan. Sebaliknya dengan memberikan jumlah *cluster* akhir yang lebih besar, maka kesamaan yang dimiliki antar kelompok akan semakin meningkat pula dan mengabaikan struktur data yang ada [7].

Salah satu metode *clustering* yang sangat populer dan banyak dipelajari untuk meminimalkan kesalahan *clustering* untuk titik ruang Euclidean disebut K-Means *clustering* [8]. Pendekatan K-Means termasuk dalam tipe partisi dalam *clustering* [8][9][10]. Pendekatan K-Means mengambil strategi *greedy* (serakah) yang menghasilkan partisi baru dengan menugaskan setiap pola ke pusat klaster terdekat dan menghitung pusat klaster baru [11]. K-Means mengklasifikasi data tertentu yang ditetapkan melalui sejumlah klaster (*k cluster*). Idenya adalah mendefinisikan nilai pusat *k* (*k centroid*), satu untuk setiap klaster. Nilai pusat ini harus ditempatkan dengan pintar karena lokasi yang berbeda menyebabkan hasil yang berbeda. Penelitian ini dilakukan untuk segmentasi pelanggan untuk membantu menganalisis data transaksi pada industri makanan, dikembangkan menggunakan metode K-Means dan model RFM. Model RFM merupakan model yang membedakan pelanggan penting dari data besar oleh tiga variabel yaitu *recency*, *frequency* dan *monetary*. Perangkat lunak yang digunakan untuk melakukan penelitian ini adalah Microsoft Excel untuk penyimpanan data dan RapidMiner sebagai perangkat lunak untuk mengevaluasi. Kombinasi antara model RFM dan metode K-Means mampu membantu dalam proses pengelompokan setiap kategori pelanggan dan mengetahui tingkat loyalitas yang dimilikinya.

## II. LANDASAN TEORI

### A. Model RFM

Model RFM (*Recency*, *Frequency* dan *Monetary*) telah banyak diterapkan dalam beberapa bidang, terutama dalam dunia pemasaran. Dengan mengadopsi model RFM, seorang pengambil keputusan dapat secara efektif mengidentifikasi pelanggan yang berharga dan akan digunakan sebagai pengembangan strategi pemasaran yang efektif [12]. Model RFM sering digunakan untuk segmentasi pasar. RFM mempertahankan informasi tentang waktu pembelian (*recency*) paling akhir, berapa kali pelanggan melakukan pembelian (*frequency*), dan rata-rata uang yang dihabiskan (*monetary*) [13]. Pelanggan yang telah membeli baru-baru ini, paling sering, dan menghabiskan banyak uang berpeluang untuk bereaksi terhadap promosi di waktu yang akan datang [14]. Keuntungan model RFM terletak pada relevansinya selama beroperasi pada beberapa variabel yang dapat diamati dan bersifat obyektif. Variabel ini

digolongkan menurut 3 kriteria, yaitu *recency*, *frequency*, dan *monetary*. *Recency* merupakan interval waktu antara pembelian terakhir dengan waktu sekarang, nilai yang lebih rendah sesuai dengan probabilitas yang lebih tinggi bahwa pelanggan akan melakukan pembelian ulang. *Frequency* adalah jumlah transaksi yang telah dilakukan pelanggan dalam jangka waktu tertentu dan *monetary* adalah jumlah uang yang dihabiskan dalam periode waktu yang ditentukan [15].

Penelitian yang dilakukan oleh Ait Daoud *et al.*, [16] menggunakan variabel *behavioral*, *recency*, *frequency* dan *monetary* yang didapatkan menggunakan model RFM, kemudian disegmentasi menggunakan metode *clustering*. Hasilnya dari 730 pelanggan terbagi menjadi 8 *cluster*, 7 *cluster* menjadi kelompok terpenting untuk penerapan strategi pemasaran di masa depan. Pada kesimpulan, penulis menuliskan bahwa model RFM dan teknik *clustering* yang diterapkan pada *online shop* dapat membantu mengidentifikasi pelanggan dengan *high-value* dan potensi keuntungan dan kemudian mendesain strategi pemasaran yang berbeda untuk setiap *cluster* yang terbentuk. Analisis RFM merupakan alat sederhana dan ampuh untuk menganalisis perilaku pelanggan. Meluasnya perdagangan melalui internet memudahkan pengumpulan riwayat pembelian dan pelacakan perilaku belanja pelanggan. Persaingan yang semakin ketat membuat pemahaman terhadap perilaku pelanggan menjadi kunci sukses sebagai masyarakat informasi [13].

Nilai *recency*, *frequency*, dan *monetary* dibagi menjadi lima bagian dengan nilai 1, 2, 3, 2, dan 5. Nilai *recency* dihitung berdasarkan tanggal transaksi terakhir atau interval waktu transaksi terakhir dengan saat ini. Nilai 5 diberikan kepada pelanggan dengan tanggal transaksi terbaru dan nilai 1 untuk pelanggan dengan tanggal transaksi terjauh di masa lalu. Kemudian *frequency*, pelanggan yang sering bertransaksi mempunyai nilai 5, sedangkan pelanggan yang jarang bertransaksi mempunyai nilai 1. Pelanggan yang membelanjakan banyak uang untuk melakukan transaksi mempunyai nilai *monetary* yang tinggi, yaitu nilai 5. Sedangkan pelanggan yang sedikit mengeluarkan uang mempunyai nilai *monetary* yang rendah, yaitu nilai 1 [17] [18]. Berikut merupakan perhitungan skor RFM:

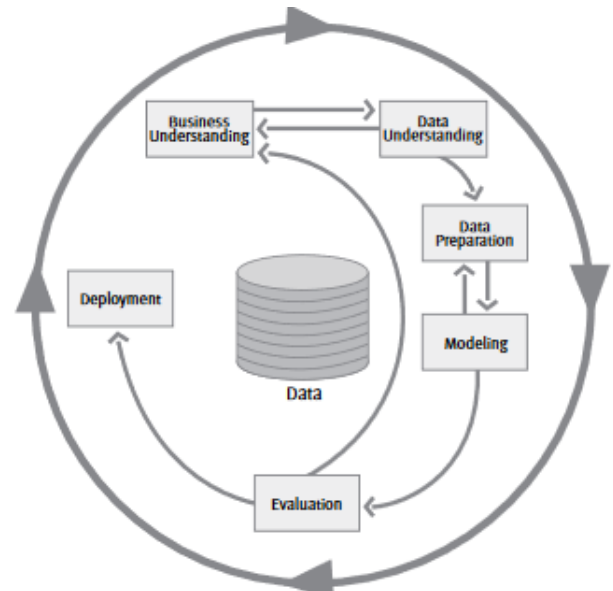
$$\begin{aligned} \text{skor RFM} = & ((\text{nilai recency} \times 100) \\ & + (\text{nilai frequency} \times 0) \\ & + \text{nilai monetary} \end{aligned} \quad (1)$$

## B. Data Mining

*Data mining* merupakan salah satu bagian langkah penting dalam proses KDD (*Knowledge Discovery in*

*Database*) terutama berkaitan dengan ekstraksi dan perhitungan pola-pola yang ditelaah.

CRISP-DM (*Cross Industry Standard Process for Data Mining*) merupakan standar proses *data mining* yang ditujukan untuk melakukan proses analisis dari suatu industri sebagai dasar strategi pemecahan masalah dari bisnis atau unit penelitian [17]. Gambar 1 menunjukkan siklus proses CRISP-DM.



Gambar 1. Proses CRISP-DM [19]

Berikut merupakan 6 fase CRISP-DM [19] :

1) *Business understanding (pemahaman bisnis)*: Tahap pertama adalah memahami tujuan dan kebutuhan dari sudut pandang bisnis, kemudian menterjemahkan pengetahuan ini ke dalam pendefinisian masalah dalam data mining. Selanjutnya akan ditentukan rencana dan strategi untuk mencapai tujuan tersebut.

2) *Data understanding (pemahaman data)*: Tahap ini dimulai dengan pengumpulan data yang kemudian akan dilanjutkan dengan proses untuk mendapatkan pemahaman yang mendalam tentang data, mengidentifikasi masalah kualitas data, atau untuk mendeteksi adanya bagian yang menarik dari data yang dapat digunakan untuk hipotesa untuk informasi yang tersembunyi.

3) *Data preparation (persiapan data)*: Tahap ini meliputi kegiatan untuk membangun kumpulan data akhir (data yang akan diproses pada tahap pemodelan) dari data mentah. Tahap ini dapat diulang beberapa kali. Pada tahap ini juga mencakup pemilihan tabel, *record*, dan atribut-atribut data, termasuk proses pembersihan dan transformasi data untuk kemudian dijadikan masukan dalam tahap pemodelan

4) *Modeling (pemodelan)*: Dalam tahap ini akan dilakukan pemilihan dan penerapan berbagai teknik pemodelan dan beberapa parameternya akan disesuaikan untuk mendapatkan nilai yang optimal. Secara khusus, ada

beberapa teknik berbeda yang dapat diterapkan untuk masalah *data mining* yang sama. Di pihak lain ada teknik pemodelan yang membutuhkan format khusus sehingga pada tahap ini masih memungkinkan kembali ke tahap sebelumnya.

5) *Evaluation (evaluasi)*: Pada tahap ini, model sudah terbentuk dan diharapkan memiliki kualitas baik jika dilihat dari sudut pandang analisis data. Pada tahap ini akan dilakukan evaluasi terhadap apakah model dapat mencapai tujuan yang ditetapkan pada fase awal (pemahaman data). Kunci dari tahap ini adalah menentukan ada atau tidak masalah bisnis yang dipertimbangkan. Pada akhir dari tahap ini harus ditentukan penggunaan hasil proses *data mining*.

6) *Deployment (pengembangan)*: Pada tahap terakhir, pengetahuan atau informasi yang telah diperoleh akan diatur dan dipresentasikan dalam bentuk khusus sehingga dapat digunakan oleh pengguna. Tahap pengembangan dapat berupa pembuatan laporan sederhana atau mengimplementasikan proses *data mining* yang berulang dalam perusahaan. Dalam banyak kasus, tahap *deployment* melibatkan konsumen, di samping analisis data, karena sangat penting bagi konsumen untuk memahami tindakan yang harus dilakukan menggunakan model yang telah dibuat.

Referensi [20] mengadopsi metodologi CRISP-DM untuk memprediksi fase pengobatan gangguan pada ibu hamil di Julio Dinis Maternity Hospital. Mereka menjelaskan proses yang rumit, ketika selesai dalam konteks metodologis, CRISP-DM menjadi lebih mudah dipahami, diterapkan dan dikembangkan.

CRISP-DM mengusulkan sebuah model proses yang komprehensif untuk melakukan proyek *data mining*. *Paper* yang ditulis oleh Rudiger dan Wilhelm [20], memberikan pendapat bahwa model proses standar akan bermanfaat bagi industri *data mining* dan menyajikan beberapa pengalaman praktis dengan metodologi CRISP-DM. Tujuannya adalah untuk menentukan suatu proses yang dapat diulang dengan andal dan efisien oleh orang yang berbeda dengan situasi yang berbeda juga. Mereka menyimpulkan bahwa CRISP-DM bekerja dengan baik. Model proses generik berguna untuk perencanaan, dokumentasi dan komunikasi. Model proses cukup mudah untuk ditulis namun tingkat ketepatan yang suliy. Proyek CRISP-DM tidak mudah untuk dievaluasi, terutama dalam hal kecepatan dan biaya. Tapi CRISP-DM mendukung proses yang berulang dan sangat berguna untuk proyek yang besar.

### C. K-Means Clustering

Teknik *data mining* telah diterapkan pada berbagai domain. Aktivitas transaksi pada sebuah organisasi yang terjadi hampir setiap hari, membuat data transaksi menjadi sangat banyak dan kompleks. Dengan teknik *data mining*, khususnya teknik *clustering*, dapat diterapkan untuk membagi seluruh pelanggan menjadi beberapa kelompok (*cluster*) berdasarkan beberapa kesamaan pada pelanggan

[21]. *Clustering* adalah studi formal untuk partisi atau mengelompokkan data dengan tidak menggunakan pelabelan kategori. *Clustering* bersifat *unsupervised learning* atau tidak mempunyai tahap pelatihan data, berbeda dengan klasifikasi. *Clustering* digunakan untuk mengelompokkan data secara alamiah berdasarkan kemiripan pada objek data dan sebaliknya meminimalkan kemiripan terhadap *cluster* lain [17].

Algoritma K-Means adalah algoritma klasik untuk memecahkan masalah *clustering* [22]. Metode K-Means merupakan metode non-hirarkis. Metode ini merupakan teknik penyekatan (*partition*) yang membagi atau memisahkan obyek k ke daerah yang terpisah [6]. Metode K-Means digunakan untuk mengelompokkan n vector berdasarkan atribut partisi k, dimana  $k < n$ , tergantung pada beberapa tindakan. Pusat *cluster* adalah *mean* (nilai tengah) semua *vector* pada *cluster* tertentu. Algoritma ini dimulai dengan memilih *centroid* k awal secara acak (*randomly*), kemudian memberikan nilai vektor ke *centroid* terdekat dengan *Euclidean distance* dan menghitung ulang *centroid* baru. Proses ini berulang sampai *vector* tidak lagi mengubah *cluster* antar iterasi [23].

Pengembangan implementasi *data mining* sangat penting untuk perusahaan telekomunikasi, yang merupakan tipe industri *data-intensive*. Segmentasi pelanggan akan membantu menganalisa komposisi pelanggan dengan akurat dan memajukan kualitas pelayanan dan pemasaran. Menggunakan K-Means *clustering* dan *data mining tool* KXEN, penelitian lainnya mengusulkan resolusi dari segmentasi pelanggan untuk Changzhou telecom di provinsi Jiangsu. Hasil menunjukkan bahwa resolusinya efektif dan sukses [5]. Kemudian penelitian berikutnya menggunakan *machine learning* untuk teknik *clustering* yang digunakan untuk segmentasi pelanggan dari sebuah distributor. Penelitian ini menjelaskan tentang seleksi, analisis dan interpretasi *cluster* untuk mengevaluasi pelanggan yang menghabiskan produk setiap tahunnya. Algoritma K-Means menjadi pilihan mereka dalam mengelompokkan pelanggan melalui perilaku pelanggan [24]. Dalam penelitian ini K-Means digunakan karena komputasi yang ringan, kemudian sesuai dengan kebutuhan data yang akan di *cluster*, dan sesuai dengan tujuan untuk menentukan jumlah *cluster* di awal untuk memperoleh jumlah kelompok pelanggan yang optimal.

### D. Segmentasi Pelanggan

Segmentasi terus menjadi konsep pemasaran yang penting juga dalam konteks *relationship marketing*. Meningkatkan hubungan dengan pelanggan menjadi lebih menarik dan akan menghasilkan pemahaman yang lebih baik tentang kebutuhan pelanggan. Paper ini [20] berpendapat bahwa analisis retrospektif atau historis yang memberikan fasilitas perhitungan profitabilitas *customer relationship*,

merupakan titik awal yang sangat baik untuk segmentasi pelanggan. Segmentasi terbukti menjadi instrumen analisis yang kuat sebagai dasar perumusan, strategi pemasaran.

Segmentasi adalah proses membagi pelanggan menjadi beberapa klaster dengan kategori loyalitas pelanggan untuk membangun strategi pemasaran. Segmentasi pelanggan adalah salah satu langkah awal dalam membuat model bisnis. Segmentasi pelanggan dibagi menjadi 6 karakteristik berdasarkan nilai RFM seperti yang ditunjukkan pada Tabel 1 [25]:

TABEL 1  
KARAKTERISTIK PELANGGAN

Kelas pelanggan	Karakteristik
<i>Superstar</i>	<ul style="list-style-type: none"> <li>a. Pelanggan dengan loyaliti yang tinggi.</li> <li>b. Mempunyai nilai <i>monetary</i> yang tinggi</li> <li>c. Mempunyai frekuensi yang tinggi</li> <li>d. Mempunyai transaksi paling tinggi</li> </ul>
<i>Golden customer</i>	<ul style="list-style-type: none"> <li>a. Mempunyai nilai <i>monetary</i> tertinggi yang ke dua</li> <li>b. <i>Frequency</i> yang tinggi</li> <li>c. Mempunyai rata-rata transaksi</li> </ul>
<i>Typical customer</i>	Mempunyai rata-rata nilai <i>monetary</i> dan rata-rata transaksi
<i>Occasional customer</i>	<ul style="list-style-type: none"> <li>a. Nilai <i>monetary</i> terendah kedua setelah <i>dormant customer</i></li> <li>b. Nilai <i>recency</i> paling rendah</li> <li>c. Transaksi paling tinggi</li> </ul>
<i>Everyday shopper</i>	<ul style="list-style-type: none"> <li>a. Memiliki peningkatan transaksi</li> <li>b. Transaksi yang rendah</li> <li>c. Mempunyai nilai <i>monetary</i> sedang sampai dengan rendah</li> </ul>
<i>Dormant customer</i>	<ul style="list-style-type: none"> <li>a. Mempunyai <i>frequency</i> dan <i>monetary</i> yang paling rendah</li> <li>b. Nilai <i>recency</i> yang paling rendah</li> </ul>

### III. METODOLOGI PENELITIAN

*Cross Industry Standard Process for Data Mining* (CRISP-DM) merupakan metode yang digunakan dalam penelitian ini. CRISP-DM adalah model proses penambangan data (*data mining*) yang sering digunakan para ahli untuk memecahkan masalah. Jalannya penelitian mengacu pada enam tahap CRISP-DM, yang meliputi proses

pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan proses yang terakhir adalah pengembangan/penyebaran.

#### A. Pemahaman Bisnis

Memahami tujuan dan kebutuhan dalam lingkup bisnis atau unit penelitian, menterjemahkan pengetahuan ini ke dalam permasalahan *data mining*. Kemudian menentukan strategi untuk mencapai tujuan tersebut. Studi kasus penelitian ini dilakukan di salah satu UKM (Usaha Kecil dan Menengah) di Kabupaten Wonosobo yaitu UD Gemilang Kencana, sebuah usaha kecil dan menengah yang memproduksi manisan buah carica. Unit dagang ini berdiri akhir tahun 2014 dengan jumlah konsumen yang terus meningkat dari tahun ke tahun. UD Gemilang Kencana melayani transaksi penjualan baik secara langsung (*offline*) dengan konsumen atau dengan media *online* seperti Whatsapp, Facebook, dan Instagram. Data transaksi yang tersimpan digunakan untuk melakukan pengelompokan atau segmentasi pelanggan UD Gemilang Kencana. Untuk mempermudah dalam proses pengelompokan pelanggan, pada penelitian ini menggunakan parameter *recency* (terakhir melakukan transaksi), *frequency* (seringnya melakukan transaksi), dan *monetary* (berapa banyak uang untuk transaksi). Selain itu untuk memudahkan dalam pengelompokan data dan pelabelan pelanggan, dalam penelitian ini menggunakan aplikasi RapidMiner untuk proses pengelompokan pelanggan secara otomatis dan dapat diketahui karakter masing-masing pelanggan.

#### B. Pemahaman data

Pada proses ini akan dilakukan pemahaman terhadap kebutuhan data terkait dengan pencapaian tujuan dalam CRM (*Customer Relationship Management*) dan *data mining*. Data diambil dari salah satu UKM di Kabupaten Wonosobo. Kemudian dilakukan pemahaman data, mengidentifikasi kualitas data, memeriksa data dan membersihkan data yang tidak valid (*data cleansing*). Penelitian ini menggunakan data transaksi penjualan manisan buah carica dari 1 Juni 2017 sampai dengan 31 Desember 2017. Data diambil dari dokumentasi UD Gemilang Kencana yang berformat xls. Jumlah data transaksi terdiri dari 127 data.

#### C. Persiapan Data

Pada tahap ini data yang tersimpan akan dipersiapkan sehingga mempermudah proses penambangan (*mining*). Ada beberapa proses yang dilakukan dapat mempersiapkan data diantaranya memilih variabel yang akan dianalisis, membersihkan data, menyiapkan data awal sehingga siap untuk *data transformation*.

Segmentasi pelanggan pada penelitian ini berdasarkan pada model RFM (*recency, frequency, monetary*), maka *data selection* dari model RFM adalah rentang waktu terakhir pelanggan melakukan transaksi dengan metode analisis, jumlah frekuensi (seberapa sering) transaksi yang dilakukan pelanggan selama periode penelitian, serta jumlah nominal transaksi untuk setiap pelanggan selama periode penelitian, yaitu selama 1 Juni sampai dengan 31 Desember 2017. Tabel 2 menunjukkan hasil dari *data selection*.

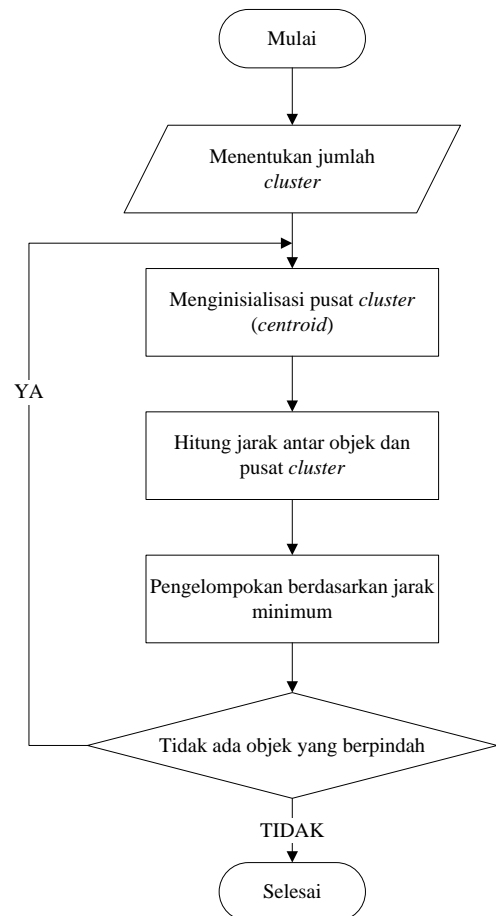
TABEL II  
HASIL SELEKSI DATA

Data Awal	Data Akhir
Tanggal terakhir transaksi (tipe: date)	<i>Recency (type: number)</i>
Jumlah melakukan transaksi	<i>Frequency</i>
Jumlah nominal transaksi	<i>Monetary</i>

Setelah melakukan *data selection*, proses selanjutnya yang harus dilakukan adalah *data preprocessing*. Proses ini berfungsi untuk memastikan kualitas data yang telah dipilih pada tahap *data selection*. Pada tahap ini diperlukan adanya pembersihan data (*data cleansing*). *Data cleansing* yang dilakukan yaitu membersihkan data RFM dari *noise* atau *missing value*, dengan membuang *noise* atau *missing value* pada data. Tahap selanjutnya adalah *data transformation* yang dilakukan dengan cara normalisasi. Normalisasi dilakukan agar skala data tidak berbeda terlalu jauh.

#### D. Pemodelan

Adapun metode yang diusulkan pada pemodelan penelitian ini adalah algoritma K-Means. Data input untuk proses *clustering* yaitu data RFM transaksi pelanggan yang sudah dinormalisasi. Proses *clustering* menggunakan algoritma K-Means ditunjukkan oleh *flowchart* pada Gambar 2.



Gambar 2. Flowchart algoritma K-Means

Berdasarkan konsep yang sudah dijelaskan, proses komputasi untuk K-Means dijelaskan sebagai berikut [26][27][28][29]:

1. Inisialisasi vektor *centroid* kluster secara acak
2. Untuk setiap data vektor, hitung jarak antar data vektor dengan masing-masing kluster *centroid* yang akan ditentukan vektor data minimum dengan kluster dan jarak yang dihitung menggunakan persamaan:

$$d(Z_p, M_j) = \sqrt{\sum_{k=1}^d (Z_{p,k} - M_{j,k})^2} \quad (2)$$

Di mana  $Z_p$  adalah titik data ke- $p$ ,  $M_j$  adalah *centroid* dari data kluster ke- $j$ .

3. Hitung ulang kluster *centroid* menggunakan persamaan:

$$M_j = \frac{1}{n_j} \left( \sum Z_p \right) \quad (3)$$

Dimana  $n_j$  adalah jumlah titik data di kluster  $j$ .

4. Ulangi langkah 3 dan 4 hingga berhenti memenuhi kriteria. Kriteria yang memuaskan dapat berupa jumlah iterasi atau perubahan posisi *centroid* dalam iterasi berurutan.

K-Means dipakai sebagai metode *clustering* pada paper ini karena kelebihananya [30], yaitu:

1. Mudah untuk diimplementasikan
2. Kemudahan dan kecepatan komputasi sehingga memungkinkan untuk digunakan pada dataset yang besar.

Selain itu K-Means juga memiliki beberapa kekurangan yang perlu diperhatikan [28][30], yaitu:

1. Pemilihan awal jumlah *cluster* (*k*) harus diketahui dan ditentukan sebelumnya oleh pengguna.
2. Hasil terakhir bergantung pada pusat *cluster* awal yang dipilih secara acak.
3. Kadang mengakibatkan adanya *cluster* yang kosong karena tidak ada titik awal yang dialokasikan saat langkah penetapan pusat *cluster* awal.

#### E. Evaluasi

Tahap evaluasi merupakan tahap interpretasi terhadap pemodelan *data mining* berdasarkan domain pengetahuan pada pemahaman bisnis. Tahap evaluasi dilakukan setelah klaster-klaster terbentuk dari hasil *clustering*. Pada tahap ini kami mengusulkan dua proses evaluasi yakni *Global Silhouette* dan *Variance*. Metode *silhouette* merupakan salah satu metode yang digunakan untuk menguji kualitas klaster yang dihasilkan dari proses *clustering*. Berikut merupakan persamaan yang digunakan pada *global silhouette* [31]:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (4)$$

Dimana:

$S(i)$  = nilai *silhouette*.

$b(i)$  = rata-rata jarak dari objek *i* dengan seluruh objek dalam *cluster* yang sama.

$a(i)$  = rata-rata jarak dari objek *i* dengan objek berada di *cluster* yang berbeda.

Nilai dari *silhouette* berkisar  $-1 \leq S(i) \leq 1$ , dimana hasil *clustering* baik jika nilai *silhouette* bernilai positif (0-1) [31]. Hal ini menandakan bahwa data berada pada *cluster* yang tepat.

Proses evaluasi yang kedua dilakukan dengan evaluasi nilai *variance* (*R*). *R* adalah nilai rasio antara jarak rata-rata data dalam satu *cluster* dan jarak rata-rata dengan *cluster* lain. Nilai *R* yang mendekati 0 menunjukkan dalam satu *cluster* memiliki kemiripan nilai yang dekat [32].

$$R = \frac{1/c \sum_{i=1}^C R_c}{1/c \sum_{i,j=1}^C R_{ij}} \quad (5)$$

Dimana:

*R* adalah nilai *variance*

*C* adalah jumlah *cluster*

$R_c$  adalah jarak rata-rata objek dalam satu *cluster*

$R_{ij}$  adalah jarak rata-rata objek dengan *cluster* lain

#### F. Pengembangan

Dalam tahap ini pengetahuan atau informasi yang telah diperoleh diimplementasikan dalam sebuah laporan agar lebih mudah dipahami oleh pihak yang berkepentingan.

### IV. HASIL DAN PEMBAHASAN

Data yang digunakan untuk pemodelan *clustering* adalah data transaksi *online* di UD Gemilang Kencana pada periode 1 Juni sampai dengan 31 Desember 2017. Data diambil secara manual yang dituliskan dalam Microsoft Excel. Tabel 3 merupakan data transaksi yang terdiri dari nama pelanggan, terakhir transaksi, frekuensi, dan jumlah uang yang dikeluarkan.

TABEL III  
CUPLIKAN DATA TRANSAKSI PELANGGAN

Nama	Terakhir transaksi	Frekuensi	Jumlah
Adi	6/5/2017	2	1,554,000
Agustinus Prie	12/23/2017	3	615,000
Aisyah Eliani	6/10/2017	3	2,205,000
Anggun Kusuma	7/15/2017	53	40,771,000
Antariksa Laksinta	10/18/2017	2	883000
Anton	12/12/2017	4	4498000
Arindro Nugroho	8/3/2017	1	761000
Astri	12/19/2017	6	5135000
Awaludin Nur Imam	10/18/2017	2	883000
Ayu	8/21/2017	4	9710000
Cahyo Eko Saputro	7/27/2017	1	1313000
Chatarin Dewi	9/13/2017	2	1320000
Danang Prayoga	7/15/2017	1	1780500
Dani Prasetyo	6/16/2017	1	1057000
Darmini	6/28/2017	1	388000
Dayat	12/25/2017	1	4338000

Setelah data disiapkan, langkah berikutnya adalah melakukan *data selection* dengan fitur *recency*, *frequency*, dan *monetary*. Kemudian dilanjutkan dengan *data transformation*. Pada *data transformation* dilakukan normalisasi. Seperti yang sudah dijelaskan, normalisasi merupakan proses transformasi dimana sebuah atribut *numeric* diskalakan dalam jarak lebih kecil Normalisasi



dilakukan agar skala data tidak terlalu jauh. Normalisasi skor  $z$  (*zero-mean normalization*) adalah salah satu metode/teknik yang diterapkan untuk normalisasi data, dimana nilai dari atribut  $A$  dinormalisasi berdasarkan rata-rata dan standar deviasi dari atribut  $A$ .  $v'$  adalah normalisasi nilai  $v$  dari atribut  $A$  [33].

$$v' = \frac{v - \bar{A}}{\sigma A} \quad (6)$$

$\bar{A}$  adalah nilai rata-rata dan  $\sigma A$  adalah standar deviasi dan atribut  $A$ .

Tabel 4 menunjukkan hasil proses normalisasi menggunakan *zero-mean normalization*.

TABEL IV  
HASIL NORMALISASI

Id_pelanggan	Zscore R	Zscore F	Zscore M
1	3.033	0.656	0.304
2	0.618	0.758	10.263
3	2.942	0.656	0.238
4	2.306	0.573	0.303
5	0.526	0.573	0.320

Normalisasi skor  $z$  dan penentuan skor RFM pada data transaksi menggunakan Microsoft Excel. Setelah mendapatkan hasil, langkah berikutnya adalah pemodelan.

Tahap pemodelan atau tahap *data mining* menggunakan metode K-Means *clustering* digunakan untuk mencapai tujuan awal penelitian yaitu melakukan pengelompokan pelanggan yang kemudian akan digunakan untuk mengetahui tingkat loyalitas pelanggan dan berhubungan dengan strategi pemasaran yang tepat untuk setiap pelanggan. K-Means sebagai metode *clustering* yang digunakan, menuntut jumlah kluster harus ditentukan oleh pengambil keputusan. *Davies bouldin index* digunakan dalam penelitian ini untuk mengidentifikasi nilai  $k$  yang optimal [33].

$$DBI = \frac{1}{k} + \sum_{i=1}^k \max_{i \neq j} (R_{i,j}) \quad (7)$$

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,j}} \quad (8)$$

$$SSB_{i,j} = d(c_i, c_j) \quad (9)$$

$$SSW = \frac{1}{m_i} + \sum_{j=i}^{m_i} d(X_j, C_i) \quad (10)$$

Nilai DBI yang kecil merupakan jumlah kluster yang baik [33]. Semakin kecil nilai DBI semakin optimal hasil kluster. RapidMiner digunakan sebagai *tool* untuk melakukan proses *clustering* dan pengujian DBI. Tabel 5

menunjukkan hasil *clustering* yang dilakukan dengan  $k=3$  menunjukkan nilai DBI yang paling baik karena menghasilkan nilai yang paling kecil [34].

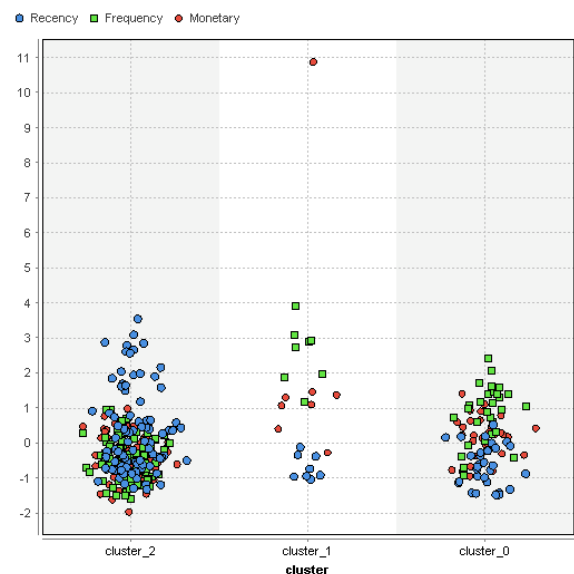
TABEL V  
HASIL PENGUJIAN DBI

Jumlah cluster (k)	Davies Bouldin Index
2	-0.879
3	-0.941
4	-0.436
5	-0.430
6	-0.530
7	-0.537
8	-0.513
9	-0.560
10	-0.602

Setelah memperoleh hasil DBI yang terbaik adalah  $k=3$  yaitu dengan nilai -0.941 dilanjutkan dengan proses *clustering*. Tabel 6 merupakan hasil pusat *cluster* yang diperoleh untuk setiap *cluster*.

TABEL VI  
HASIL PUSAT CLUSTER

Cluster	Pusat R	Pusat F	Pusat M
C-1	0.604	0.844	0.199
C-2	0.706	2.743	2.282
C-3	0.267	0.531	0.272



Gambar 3. Chart hasil proses *clustering* dengan atribut RFM

Gambar 3 menunjukkan titik persebaran setiap data yang dihasilkan dari proses *clustering*. Setiap *cluster* memiliki karakteristik yang berbeda-beda. Titik berwarna biru menunjukkan variabel *monetary*, titik hijau adalah variabel *frequency* dan titik merah untuk variabel *monetary*.

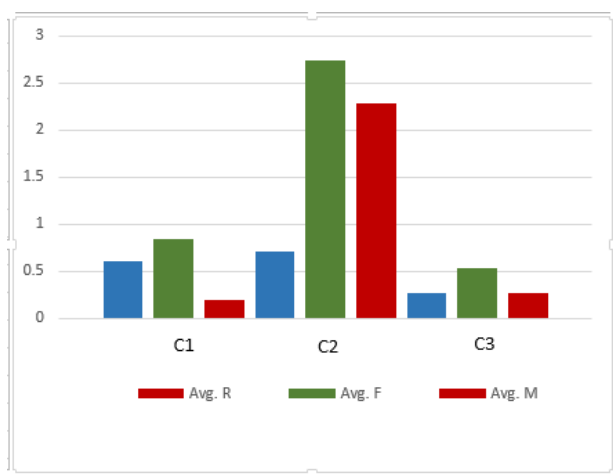


*Cluster 3 (cluster\_2)* menjadi kelompok yang memiliki titik data terbanyak dibanding dua *cluster* lainnya. Kemudian titik warna dalam *cluster 3* terlihat berdekatan karena adanya kemiripan nilai RFM. Hasil proses *clustering* ditunjukkan pada Tabel 7.

TABEL VII  
HASIL SEGMENTASI PELANGGAN BERDASARKAN RFM

Cluster	Jumlah pelanggan	Rata-rata R	Rata-rata F	Rata-rata M	Skor RFM
C-1	30	0.60432	0.84352	0.19899	245-341
C-2	8	0.70618	2.743035	2.2822249	432-545
C-3	89	0.267178	0.53093	0.27222	111-231

Dari 127 data pelanggan, setelah melalui proses *clustering* diperoleh tiga *cluster* dengan karakteristiknya masing-masing. Pengujian jumlah *cluster* dilakukan menggunakan *davies bouldin index* untuk hasil yang optimal. *Cluster* pertama terdiri dari 30 pelanggan yang termasuk dalam kelompok *typical customer* dengan skor RFM antara 245-341. *Cluster* kedua terdiri dari 8 pelanggan yang termasuk dalam kelompok *superstar* dengan skor RFM antara 432-545 dan *cluster* tiga merupakan kelompok dengan *value* terendah yaitu antara 111-231 yang masuk ke dalam kategori *dormant customer*.



Gambar 4. Grafik Hasil Clustering

Pada Gambar 4 menunjukkan bahwa *cluster 2 (C2)* merupakan *cluster* dengan nilai RFM tertinggi. Pelanggan pada *cluster* ini adalah pelanggan yang paling sering melakukan transaksi dengan jumlah uang yang dikeluarkan paling banyak dibandingkan dua *cluster* lainnya. *Cluster 3* merupakan kelompok dengan nilai RFM terendah. Pelanggan pada *cluster* ini 1-2 kali melakukan transaksi dengan jumlah uang yang dikeluarkan tidak banyak.

Kemudian grafik *cluster 1* memiliki nilai R dan F di atas *cluster 3*, namun dengan nilai M di bawah *cluster 3*.

Setelah memperoleh hasil *cluster* melalui proses *clustering* dengan algoritma K-Means, tahap berikutnya adalah melakukan pengujian atau evaluasi. Tahap ini bertujuan untuk menguji hasil *cluster* sudah optimal atau belum. Pengujian dilakukan sebanyak dua kali, yaitu dengan *global silhouette* dan *variance*.

Pengujian *silhouette* dilakukan terhadap data transaksi yang sudah dikumpulkan, yaitu periode 1 Juni sampai dengan 31 Desember 2017. Pengujian dilakukan dengan masukan tiga *cluster* dan dihitung menggunakan persamaan *silhouette*, hasilnya sebesar 0.2167. Hasil menunjukkan nilai positif, sehingga dapat dikatakan proses *clustering* berjalan dengan baik. Dapat disimpulkan bahwa hasil yang positif menunjukkan bahwa jarak rata-rata antar objek dalam *cluster* yang sama ( $b(i)$ ) lebih kecil dibanding jarak rata-rata antar objek dengan objek yang ada di *cluster* lain ( $a(i)$ ). Masing-masing pelanggan berada pada *cluster* yang tepat.

Pengujian kedua dengan evaluasi *variance* yang digunakan untuk menguji kualitas *cluster* yang terbentuk. *Cluster* yang baik adalah *cluster* yang memiliki nilai jarak mendekati 0 dalam *cluster*. Hal ini menunjukkan kemiripan nilai data yang dekat. Pada proses pembentukan tiga *cluster*, menunjukkan hasil evaluasi *variance* sebesar 1.326. Nilai *variance* menunjukkan angka yang mendekati 0. Sehingga dapat disimpulkan pembentukan tiga *cluster* memiliki kemiripan nilai data yang dekat untuk anggota setiap *cluster*. Hasil dua pengujian tersebut ditunjukkan pada Tabel 8.

TABEL VIII  
HASIL PENGUJIAN

Jumlah cluster (n)	Evaluasi	
	Global Silhouette	Variance
3	0.2167	1.326

## V. KESIMPULAN

Tujuan utama penelitian ini adalah segmentasi pelanggan dan mengukur kesetiaan mereka terhadap sebuah produk UKM dengan mengkombinasikan model RFM dan teknik *clustering* menggunakan metode K-Means. Model RFM yang terdiri dari *recency*, *frequency* dan *monetary* digunakan sebagai variabel/atribut data yang digunakan untuk proses *clustering*. Untuk menemukan jumlah *cluster* yang paling optimal dilakukan perhitungan menggunakan *davies bouldin index*. Hasilnya terdapat 3 *cluster* dengan karakteristiknya masing-masing. *Cluster* pertama dengan 30 pelanggan masuk ke dalam kategori *typical customers*. *Cluster* kedua terdiri dari 8 pelanggan masuk ke dalam kategori *superstar*. Kemudian 89 pelanggan termasuk dalam kategori *dormant customer*. Kemudian pengujian dilakukan menggunakan *global silhouette* dan *variance*. Hasilnya menunjukkan bahwa data di setiap *cluster* berada

pada *cluster* yang tepat dan memiliki kemiripan yang dekat. Aplikasi model RFM dan teknik *clustering* khususnya algoritma K-Means, dapat membantu mengidentifikasi pelanggan dengan nilai tinggi dan berpotensi memberi keuntungan serta dapat membantu dalam merancang strategi pemasaran yang tepat untuk ketiga *cluster* yang terbentuk.

#### DAFTAR PUSTAKA

- [1] H. Zhao and C. He, "Objective cluster analysis in value-based customer segmentation method," *Proc. - 2009 2nd Int. Work. Knowl. Discov. Data Mining, WKKD 2009*, pp. 484–487, 2009.
- [2] W. Bi, M. Cai, M. Liu, and G. Li, "A Big Data Clustering Algorithm for Mitigating the Risk of Customer Churn," *IEEE Trans. Ind. Informatics*, vol. 12, no. 3, pp. 1270–1281, 2016.
- [3] Y. Chen, G. Zhang, D. Hu, and S. Wang, "Customer Segmentation in Customer Relationship Management Based on Data Mining," vol. 207, pp. 288–293, 2006.
- [4] I. Soesanti, "Web-Based Monitoring System on the Production Process of Yogyakarta Batik Industry," *J. Theor. Appl. Inf. Technol.*, vol. 87, no. 1, pp. 146–152, 2016.
- [5] Y. Luo, Q. R. Cai, H. X. Xi, Y. J. Liu, and Z. M. Yu, "Telecom customer segmentation with K-means clustering," *ICCSE 2012 - Proc. 2012 7th Int. Conf. Comput. Sci. Educ.*, no. Iccse, pp. 648–651, 2012.
- [6] D. Zheng, "Application of silence customer segmentation in securities industry based on fuzzy cluster algorithm," *J. Inf. Comput. Sci.*, vol. 10, no. 13, pp. 4337–4347, 2013.
- [7] R. J. Kuo, S. H. Lin, and Z.-Y. Chen, "Integration of Particle Swarm Optimization and Immune Genetic Algorithm-Based Dynamic Clustering for Customer Clustering," *Int. J. Artif. Intell. Tools*, vol. 24, no. 5, p. 1550019, 2015.
- [8] W. Li, "Modified K-Means Clustering Algorithm," *2008 Congr. Image Signal Process.*, pp. 618–621, 2008.
- [9] N. Kurinjivendhan and K. Thangadurai, "Modified k-means algorithm and genetic approach for cluster optimization," *Proc. 2016 Int. Conf. Data Min. Adv. Comput. SAPIENCE 2016*, pp. 53–56, 2016.
- [10] M. K. Algorithm and B. D. Clustering, "IBAIS University," 2015.
- [11] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, 2010.
- [12] J. Wei, S. Lin, and H. Wu, "A review of the application of RFM model," *African J. Bus. Manag.*, vol. 4, no. 19, pp. 4199–4206, 2010.
- [13] Y. J. Lee, "Privacy-preserving Data Mining for Personalized Marketing," *Int. J. Comput. Commun. Networks*, vol. 4, no. 1, pp. 1–9, 2014.
- [14] A. X. Yang, "How to develop new approaches to RFM segmentation," *J. Targeting, Meas. Anal. Mark.*, vol. 13, no. 1, pp. 50–60, 2004.
- [15] C. Wang, "Robust Segmentation for the Service Industry Using Kernel Induced Fuzzy Clustering Techniques," *Proc. 2009 IEEE IEEM*, pp. 2197–2201, 2009.
- [16] R. Ait Daoud, A. Amine, B. Bouikhalene, and R. Lbibb, "Combining RFM model and clustering techniques for customer value analysis of a company selling online," *Proc. IEEE/ACS Int. Conf. Comput. Syst. Appl. AICCSA*, vol. 2016–July, 2016.
- [17] C. Cheng and Y. Chen, "Expert Systems with Applications Classifying the segmentation of customer value via RFM model and RS theory," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 4176–4184, 2009.
- [18] S. M. S. Hosseini, A. Maleki, and M. R. Gholamian, "Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty," *Expert Syst. Appl.*, vol. 37, no. 7, pp. 5259–5264, 2010.
- [19] P. Chapman *et al.*, "Crisp-Dm 1.0," *Cris. Consort.*, p. 76, 2000.
- [20] A. Brandão, E. Pereira, F. Portela, M. F. Santos, A. Abelha, and J. Machado, "Managing Voluntary Interruption of Pregnancy Using Data Mining," *Procedia Technol.*, vol. 16, pp. 1297–1306, 2014.
- [21] S.-C. Huang, E.-C. Chang, and H.-H. Wu, "A case study of applying data mining techniques in an outfitter's customer value analysis," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 5909–5915, 2009.
- [22] J. Zhao, W. Zhang, and Y. Liu, "Improved K-Means cluster algorithm in telecommunications enterprises customer segmentation," *Proc. 2010 IEEE Int. Conf. Inf. Theory Inf. Secur. ICITIS 2010*, pp. 167–169, 2010.
- [23] D. Birant, "Data Mining Using RFM Analysis," in *Knowledge-Oriented Applications in Data Mining*, Turkey: KImto Funatso, 2011, pp. 91–108.
- [24] I. Pranata and G. Skinner, "Segmenting and targeting customers through clusters selection & analysis," *ICACIS 2015 - 2015 Int. Conf. Adv. Comput. Sci. Inf. Syst. Proc.*, pp. 303–308, 2016.
- [25] K. Tsipstsis, *Data Mining Tehniques in CRM: Inside Customer Segmentation*. 2010.
- [26] S. Singh Raghuwanshi and P. Arya, "Comparison of K-means and Modified K-mean algorithms for Large Data-set," *Int. J. Comput. Commun. Netw.*, vol. 1, no. 3, pp. 106–110, 2012.
- [27] R. V. Singh and M. P. S. Bhatia, "Data clustering with modified K-means algorithm," *Int. Conf. Recent Trends Inf. Technol. ICRITIT 2011*, pp. 717–721, 2011.
- [28] B. Yi, F. Yang, H. Qiao, and C. Xu, "An improved initialization center algorithm for K-means clustering," *2010 Int. Conf. Comput. Intell. Softw. Eng. CiSE 2010*, no. 1, pp. 1–4, 2010.
- [29] T. Widiyaningtyas, M. Indra, W. Prabowo, and M. A. M. Pratama, "Implementation of K-Means Clustering Method to Distribution of High School Teachers," no. September, pp. 19–21, 2017.
- [30] K. Singh, D. Malik, and N. Sharma, "Evolving limitations in K-means algorithm in data mining and their removal," *IJCEM Int. J. Comput. Eng. Manag. ISSN*, vol. 12, no. April, pp. 2230–7893, 2011.
- [31] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. C, pp. 53–65, 1987.
- [32] T. Handhayani, I. Wasito, M. Sadikin, and A. I. K-means, "Kernel Based Integration of Gene Expression and DNA Copy Number," pp. 978–979, 2013.
- [33] T. Hardiani, S. Sulistyono, and R. Hartanto, "Segmentasi Nasabah Tabungan Menggunakan Model RFM ( Recency , Frequency , Monetary ) dan K-Means Pada Lembaga Keuangan Mikro ISBN : 979-26-0280-1 ISBN : 979-26-0280-1," *Semin. Nas. Teknol. Inf. dan Komun. Terap.*, no. May 2017, pp. 463–468, 2015.
- [34] H. Qiao and B. Edwards, "A data clustering tool with cluster validity indices," *ICC2009 - Int. Conf. Comput. Eng. Sci. Inf.*, vol. 1, no. 2, pp. 303–309, 2009.